An informational distance for estimating the faithfulness of a possibility distribution, viewed as a family of probability distributions, with respect to data

Mathieu Serrurier, Henri Prade

An informational distance for estimating the faithfulness of a possibility distribution, viewed as a family of probability distributions, with respect to data

Mathieu Serrurier *, Henri Prade

UPS, IRIT, 118 Route de Narbonne, Toulouse, France

Keywords:
Possibility theory
Probability–possibility transformation
Informational distance

ABSTRACT

An acknowledged interpretation of possibility distributions in quantitative possibility theory is in terms of families of probabilities that are upper and lower bounded by the associated possibility and necessity measures. This paper proposes an informational distance function for possibility distributions that agrees with the above-mentioned view of possibility theory in the continuous and in the discrete cases. Especially, we show that, given a set of data following a probability distribution, the optimal possibility distribution with respect to our informational distance is the distribution obtained as the result of the probability–possibility transformation that agrees with the maximal specificity principle. It is also shown that when the optimal distribution is not available due to representation bias, maximizing this possibilistic informational distance provides more faithful results than approximating the probability distribution and then applying the probability–possibility transformation. We show that maximizing the possibilistic informational distance is equivalent to minimizing the squared distance to the unknown optimal possibility distribution. Two advantages of the proposed informational distance function is that (i) it does not require the knowledge of the shape of the probability distribution that underlies the data, and (ii) it amounts to sum up the elementary terms corresponding to the informational distance between the considered possibility distribution and each piece of data. We detail the particular case of triangular and trapezoidal possibility distributions and we show that any unimodal unknown probability distribution can be faithfully upper approximated by a triangular distribution obtained by optimizing the possibilistic informational distance.

1. Introduction

Possibility theory, based on max-decomposable set-functions, associated with possibility distributions, may have either a qualitative or a quantitative understanding [6], depending on the nature of the scale used for possibility degrees. Quantitative possibility theory corresponds to the case where the interval $[0, 1]$ is used as a ratio scale. In qualitative possibility theory, only the ordering of the possibility values makes sense. In this paper, we deal with quantitative possibility theory. Quantitative possibility measures can be viewed as upper bound of probabilities. Then, a possibility distribution represents a family of probability distributions [4]. The quantitative view was first suggested by Zadeh [17] when he expressed the idea that what is probable must be possible. Following this intuition a probability–possibility transformation has been proposed [7]. This transformation associates a probability distribution with the maximally specific possibility distribution which is such that the possibility of any event is indeed an upper bound of the corresponding probability.

* Corresponding author.
E-mail address: mathieu.serrurier@gmail.com (M. Serrurier)

We call informational distance any function that evaluates the adequateness between a distribution of probability (or possibility) and a set of data. Thus, the likelihood is a well-known informational distance used for building a probability distribution from a set of data. Given a set of parametrized probability distributions and a set of data, the likelihood function is used for optimizing the choice of the parameters in order to determine the best suited distribution with respect to the data. However, this approach supposes that the shape of the distribution is a priori known. Otherwise, generic probability distributions, such as Gaussian mixtures, can be used. However, it requires to have a large amount of data at our disposal. Moreover, due to the constraints on the probability distributions, induced by their additive nature, there are no simple distribution that allows for a faithful description of any set of data. An informational distance function is also useful for comparing the relative adequateness of two distributions with respect to the same set of data.

There exist different kinds of methods for eliciting possibility distributions from data. For instance, some approaches build directly the possibility distribution on the basis of a proximity relation defined on the universe of the data [5]. This is to be related to the idea of building fuzzy histograms based on fuzzy partitions, which has been fully investigated in [16,10]. Mauris proposes a method for constructing a possibility distribution when only very few data are available (even only one or two) based on probability inequalities [12,13]. This latter method is justified in the probabilistic view of possibility theory. In the same setting, Aregui and Denoeux studies the building of possibility distributions that bound a family of probability distributions on the basis of the confidence intervals of the parameters of these latter distributions [1]. This kind of approach has also been studied in the discrete case by using simultaneous confidence intervals [11]. These methods, how different they are, have in common to build the distributions directly. In this paper, we investigate another road based on the optimization of an appropriate informational distance function. The proposed informational distance function is in agreement with the view underlying the probability–possibility transformation, as we shall see. This paper is a fully revised and expanded version of [15], which is also the basis of the regression method at work in [14]. More precisely, the present paper proposes an interpretation of the possibilitstic informational distance function and establishes different properties that formally justify it and supports its practical use.

The paper is organized as follows. Section 2 provides the necessary background about possibility distributions and their interpretations in terms of families of probabilities. Section 3 presents informational distance functions for probability distributions, and in particular a new non-logarithmic one, which can be used for approximating unbounded distributions by bounded ones. Sections 4 and 5 focus on the definition of the possibilistic counterpart of the new informational distance function, in the discrete and continuous settings respectively. The specific cases of triangular or trapezoidal distributions are also discussed in this section. We propose some examples of the construction of a possibility distribution from data using the possibilistic informational distance function in Section 6. Lastly, we emphasize the usability of this function in the conclusion.

2. Possibility theory: basic settings

In this background section, we first recall the basic definitions of possibility theory. We then present the view of a possibility measure as representing a family of probabilities, and the probability–possibility transformation.

2.1. Basic notions

Possibility theory, introduced by Zadeh [17], was initially created in order to deal with imprecision and uncertainty due to incomplete information as the one provided by linguistic statements. This kind of epistemic uncertainty cannot be handled by a single probability distribution, especially when a priori knowledge about the nature of the probability distribution is lacking. A possibility distribution π is a mapping from Ω to $[0, 1]$ (Ω may be a discrete universe, i.e., $\Omega = \{C_1, \ldots, C_q\}$, or a continuous one, i.e., $\Omega = \mathbb{R}$). The value $\pi(x)$ is called the possibility degree of the value x in Ω. For any subset of Ω, the possibility measure is defined as follows:

$$\forall E \subseteq \Omega,\ \Pi(E) = \sup_{x \in E} \pi(x).$$

If it exists at least a value $x \in \Omega$ for which we have $\pi(x) = 1$, the distribution is normalized. We can distinguish two extreme cases of information situations:

- complete knowledge: $\exists x \in \Omega$ such as $\pi(x) = 1$ and $\forall y \in \Omega, y \neq x, \pi(y) = 0$
- total ignorance: $\forall x \in \Omega, \pi(x) = 1$.

The necessity is the dual measure of the possibility measure. We have:

$$\forall E \subseteq \Omega, N(E) = 1 - \Pi(\bar{E}).$$

Let us introduce the α-cuts of π. They are subsets of Ω such that:

$$A_\alpha = \{x \in \Omega, \pi(x) \geq \alpha\}.$$

Then, it can be checked that, if $\Omega = \mathbb{R}$ and the distribution is continuous and normalized, we have $\forall \alpha \in [0, 1]$, $\Pi(A_\alpha) = 1$ and $N(A_\alpha) = 1 - \alpha$.

Lastly, a possibility distribution π_1 is said to be more specific than a possibility distribution π_2 ($\pi_1 \preceq \pi_2$) if and only if

$$\forall x \in \Omega, \pi_1(x) \leq \pi_2(x)$$

2.2. A possibility distribution as a family of probability distributions

One view of possibility theory is to consider a possibility distribution as a family of probability distributions (see [2] for an overview). Thus, a possibility distribution π will represent the family of the probability distributions for which the measure of each subset of Ω will be respectively lower and upper bounded by its necessity and its possibility measures. More formally, if $\mathcal{P}$ is the set of all probability distributions defined on Ω, the family of probability distributions $\mathcal{P}_\pi$ associated with π is defined as follows:

$$\mathcal{P}_\pi = \{p \in \mathcal{P}, \forall E \subseteq \Omega, N(E) \leq P(E) \leq \Pi(E)\}. \tag{1}$$

where P is the probability measure associated with p. In this scope, the situation of total ignorance corresponds to the case where all probability distributions are possible. This type of ignorance cannot be described by a single probability distribution.

When $\Omega = \mathbb{R}$, this family of probability distributions can also be described in terms of confidence intervals. Given a probability distribution p, a confidence interval I_α is a subset of Ω such as $P(I_\alpha) = \alpha$. We define I^*_α as the smallest confidence interval with probability measure equal to α. In the following, we will only use the expression 'confidence interval' for referring to I^*_α. It can be observed that:

$$\forall \alpha, \exists \beta, I^*_\alpha = \{x | p(x) \geq \beta\}.$$

Moreover, if the distribution has a finite number of modes, I^*_α is a finite union of intervals. When π is continuous, we have:

$$\forall p \in \mathcal{P}_\pi, \forall I^*_\alpha \in \Omega, I^*_\alpha \subseteq A_{1-\alpha} \tag{2}$$

where $A_{1-\alpha}$ is the $(1 - \alpha)$-cut of π. Indeed we have $\forall E \subseteq \Omega, N(E) \leq \alpha \Rightarrow E \subseteq A_{1-\alpha}$ and $N\left(I^*_\alpha\right) \leq \alpha$ (Eq. (1)). Thus, the α confidence interval of a probability distribution in $\mathcal{P}_\pi$ is bounded by the $(1 - \alpha)$-cut of π. In this scope, a possibility distribution can be viewed as an upper bound of the confidence interval of a family of probability distributions.

2.3. Probability–possibility transformation

According to this probabilistic interpretation, a method for transforming probability distributions into possibility distributions has been proposed in [7]. The idea behind this is to choose the most informative possibility measure that upper bounds the considered probability measure. This possibility measure corresponds to the most specific possibility distribution which bounds the distribution. We denote π^{sp} the probability–possibility transformation of p. This distribution is defined in the following way:

$$\forall \pi, p \in \mathcal{P}_\pi \Rightarrow \pi^{sp} \preceq \pi. \tag{3}$$

and is obtained with this equation:

$$\forall x \in \Omega, \pi^{sp}(x) = \max_{\alpha, x \in I_\alpha} (1 - \alpha). \tag{4}$$

Let us first consider the discrete case ($\Omega = \{C_1, \ldots, C_q\}$) and a set of data $X = \{x_1, \ldots, x_n\}$ that are realizations of a random variable on Ω. Let $\alpha_1, \ldots, \alpha_q$ be the frequency of the elements of X that belong respectively to $\{C_1, \ldots, C_q\}$. Let us also assume that the frequencies of examples in class C_i are put in decreasing order, i.e., $\alpha_1 \geq \cdots \geq \alpha_q$. In the following, given a possibility distribution π, we note π_i the value $\pi(x \in C_i)$ (or $\pi(C_i)$). It has been shown in [3] that the transformation of p (which is derived from the frequencies) into a possibility distribution π^{sp} (see Eq. (4)), is:

$$\forall i \in \{1, \ldots, q\}, \pi^{sp}_i = \sum_{j=i}^{q} \alpha_j. \tag{5}$$

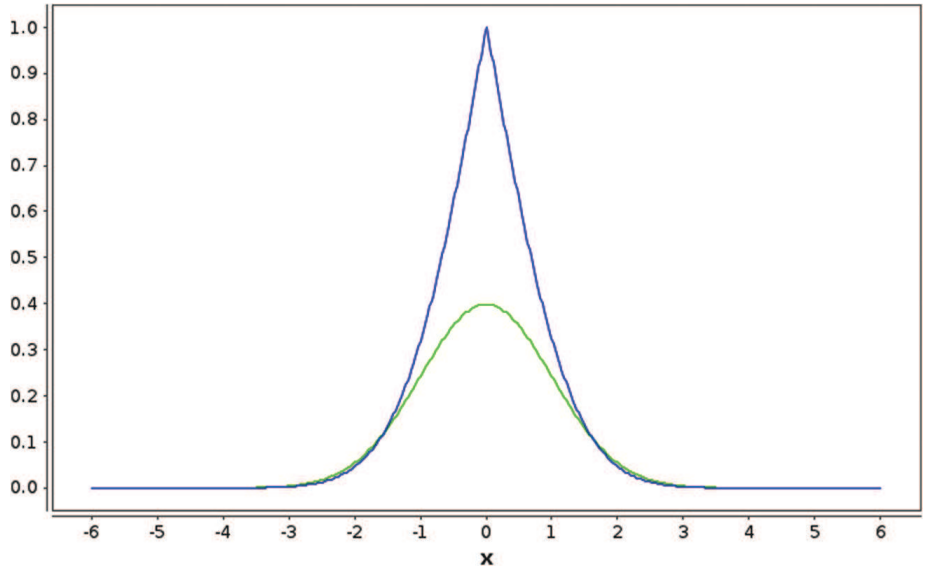

Fig. 1. Probability to possibility transformation of a Gaussian distribution.

This possibility distribution is one of the cumulated functions of p. It is worth noticing that it is the tightest one.

Example 1. For instance, we consider X that leads to the frequency $\alpha_1 = 0.5, \alpha_2 = 0.3, \alpha_3 = 0.2$. We obtain $\pi_1^{sp} = 0.5 + 0.3 + 0.2 = 1, \pi_2^{sp} = 0.3 + 0.2 = 0.5$ and $\pi_3^{sp} = 0.2$.

In the case where Ω is continuous ($\Omega = \mathbb{R}$) given p and its transformed distribution π^{sp} we have:

$$A^*_{1-\alpha} = I^*_\alpha$$

where $A^*_{1-\alpha}$ is the $(1-\alpha)$-cut of π^{sp}. Thus, if p has a finite number of modes, π^{sp} is the possibility distribution for which each $(1-\alpha)$-cuts correspond to the confidence interval of p. When p is unimodal, the unique value x such that $\pi^{sp}(x) = 1$ is the mode of p. Fig. 1 illustrates this transformation for a Gaussian distribution. The density function is in green, its possibility transformation in blue. The α-cuts of the possibility distribution corresponds to the $(1-\alpha)$ confidence interval of the probability distribution.

3. Probabilistic informational distance functions

Informational distance function is commonly used for evaluating the adequateness of a probability distribution with respect to a set of data. In the following, the likelihood function is first recalled as a noticeable informational distance function. Then, we shall look for an informational distance of the form $\mathcal{I}(f, X)$ where f is a distribution (probabilistic or possibilistic) and $X = \{x_1, \ldots, x_n\}$ is a set of data, which is decomposable, as the likelihood function, i.e.,

$$\mathcal{I}(f, X) = \sum_{i=1}^{n} \mathcal{I}(f, x_i) \tag{6}$$

Let us consider a set of realizations $X = \{x_1, \ldots, x_n\}$ of a random variable over a discrete universe $\Omega = \{C_1, \ldots, C_q\}$. We note $\alpha_1, \ldots, \alpha_q$ the frequency of the elements of X that belong respectively to $\{C_1, \ldots, C_q\}$. Given a probability distribution p on the discrete space $\Omega = \{C_1, \ldots, C_q\}$, we define $p_1, \ldots, p_q$ the probability of belonging to each element of Ω, i.e., $p(x \in C_i) = p_i$. The values $p_i, \ldots, p_q$ entirely define p, and are then the parameters of p. The maximization of the informational distances is used for estimating the parameters of a probability distribution with respect to the data. In the continuous case, the shape of the distribution has to be known, and the parameters are obtained through an optimization procedure with respect to the informational distances function. In the discrete case, the parameters are $p_i, \ldots, p_q$, and obey the constraint $\sum_i^q p_i = 1$.

The likelihood is the most used informational distance function for probability distribution. Formally the likelihood coincides to a probability value. The logarithmic-based likelihood is defined as follows (under the strict constraint $\sum_i^q p_i = 1$):

$$\mathcal{I}_{log}(p|X) = -\sum_{i=1}^{n} log(p(x_i))$$

or, when considering frequency directly

$$\mathcal{I}_{log}(p|X) = -\sum_{i=1}^{q} \alpha_i log(p_i).$$

It is equivalent to compute the joint probability of the elements of x with respect to p. As an informational distance, the likelihood has a strong limitation, since it gives a very high weight to the error when probability is very low. This is especially true when Ω is continuous. Since $\mathcal{I}_{log}$ is not defined when $p(x_i) = 0$, an unbounded density cannot be approximated by a bounded one by optimization of $\mathcal{I}_{log}$. In order to overcome these limitations, we propose another informational distance function, named $\mathcal{I}_{surf}$, that is based on the distance between the probability distribution considered and the optimal one. We have:

$$\mathcal{I}_{surf}(p|X) = \left(\sum_{i=1}^{n} p(x_i)\right) - \frac{1}{2} * \sum_{i=1}^{q} p_i^2$$

or, when considering frequency directly

$$\mathcal{I}_{surf}(p|X) = \left(\sum_{i=1}^{q} \alpha_i * p_i\right) - \frac{1}{2} * \sum_{i=1}^{q} p_i^2.$$

Roughly speaking, $\mathcal{I}_{surf}$ favors the probability distributions that share the maximum surface with the optimal one. Thus, when, Ω is continuous, it allows the approximation of unbounded densities by bounded ones.

Proposition 1. *Given a set of realization $X = \{x_1, \ldots, x_n\}$ of a random variable over a discrete universe $\Omega = \{C_1, \ldots, C_q\}$, finding the distribution p that maximizes $\mathcal{I}_{surf}$ is equivalent to find the distribution that minimizes the square of the distance to the optimal one p^*.*

Proof. Given $X = \{x_1, \ldots, x_n\}$, the optimal p^* is $p_i^* = \alpha_i$

$$\begin{aligned} d(p^*, p)^2 &= d(\alpha, p)^2 \\ &= \sum_{i=1}^{q} (\alpha_i - p_i)^2 \\ &= \sum_{i=1}^{q} \alpha_i^2 - 2 * \left(\sum_{i=1}^{q} \alpha_i * p_i - \frac{1}{2} \sum_{i=1}^{q} p_i^2\right) \\ &= C_p - 2 * \mathcal{I}_{surf}(\alpha, p) \end{aligned}$$

where $C_p = \sum_{i=1}^{q} \alpha_i^2$ is independent of p. □

Proposition 2. *Given a set of data $X = \{x_1, \ldots, x_n\}$ belonging to a discrete universe $\Omega = \{C_1, \ldots, C_q\}$, we have*

$$\underset{p \in \mathcal{P}}{argmax}(\mathcal{I}_{log}(p|X)) = \underset{p \in \mathcal{P}}{argmax}(\mathcal{I}_{surf}(p|X)).$$

Proof. Let $p_{log} = argmax_{p \in \mathcal{P}}(\mathcal{I}_{log}(p|X))$ be the optimal probability distribution given X. This distribution is such as the probability of an event C_i is equal to the frequency of element of X in C_i, i.e., $p_{log}(x \in C_i) = p_i = \alpha_i$. We now look for the probability distribution p_{surf} that maximizes $\mathcal{I}_{surf}$. We have:

$$\forall i = 1 \ldots q, \frac{\delta \mathcal{I}_{surf}(p|X)}{\delta p_i} = \alpha_i - p_i$$

thus

$$\forall i = 1 \ldots q, \frac{\delta \mathcal{I}_{surf}(p|X)}{\delta p_i} = 0 \Leftrightarrow p_i = \alpha_i.$$

Since the derivative of $\mathcal{I}_{surf}(p|X)$ with respect to p_i (the parameters of p) does not depend on the other parameters $p_j, j \neq i$, we obtain $p_{surf}(x \in C_i) = p_i = \alpha_i$. Thus $p_{surf} = p_{log}$. □

This proposition proves that, given X, the probability distribution that maximizes $\mathcal{I}_{log}$ is the same as the one that maximizes $\mathcal{I}_{surf}$.

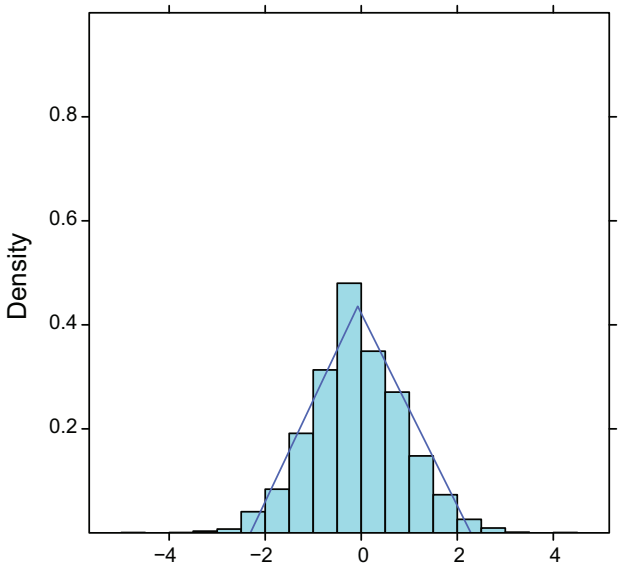

Fig. 2. Approximation of a Gaussian density with a triangular density by maximization of $\mathcal{I}_{surf}$.

$\mathcal{I}_{surf}$ can be extended to the case of $\Omega = \mathbb{R}$ in the following way:

$$\mathcal{I}_{surf}(p|X) = \left(\sum_{i=1}^{n} p(x_i)\right) - \frac{1}{2} * \int_{\mathbb{R}} p(t)^2 dt.$$

Proposition 2 remains true when $\Omega = \mathbb{R}$. Fig. 2 shows the triangular density that maximizes $\mathcal{I}_{surf}$ given a set of 3000 values that are generated by a Gaussian distribution with mean equal to 0 and standard derivation equal to 1. This triangular probability distribution is parametrized by m, l and r which are respectively the mode, the left and the right spread of the probability distribution. We have used an optimization procedure in order to find the parameters that maximizes $\mathcal{I}_{surf}$. What we obtain is the triangular probability distribution that shares the maximum of surface with the histogram computed from the sample set. If we use $\mathcal{I}_{log}$, the support of the triangle would grow to infinity as the number of values increases.

4. Possibilistic informational distance: discrete case

In this section we show how to use $\mathcal{I}_{surf}$ in order to define an informational distance function for possibility distributions that supports the interpretation of a possibility distribution in terms of a family of probability distributions. Roughly speaking, the idea is to have an informational distance function that is maximal for the result of the probability–possibility transformation of the distribution and that favors the approximations of the distribution that share the maximal surface with the optimal one (as for $\mathcal{I}_{surf}$ in the probabilistic case). We first consider the case of a discrete universe, i.e., $\Omega = \{C_1, \ldots, C_q\}$. In this case, it is expected that the possibility distribution π^* that maximizes the informational distance is exactly π^{sp} defined in Eq. (4). This possibility distribution is one of the cumulated functions of p^*. What we want to obtain is a function that is maximized for a possibility distribution π having following properties:

- π is a cumulated function of p
- $\forall i, j, \pi_i \geq \pi_j \Leftrightarrow p_i \geq p_j$.

In a first step, we will use $\mathcal{I}_{surf}$ for estimating the possibility distribution π as the cumulated function that respects the ordering of the possibility degrees in π. Then, in the following we consider that the ordering of subsets C_i and the frequencies α_i reflect the possibility ordering, i.e., $\pi_1 \geq \cdots \geq \pi_q$. By considering π as a cumulated function, we suppose that, for all i, the pair $(\pi_i, 1 - \pi_i)$ is a binomial probability distribution for the sets $\bigcup_{j=i}^{q} C_j$ and $\bigcup_{j=1}^{i-1} C_j$. By estimating this probability function by $\mathcal{I}_{surf}$ we obtain:

$$\begin{aligned}\mathcal{I}_{surf}((\pi_i, 1-\pi_i)|X) &= \left(\sum_{j=i}^{q} \alpha_j\right) * \pi_i + \left(\sum_{j=1}^{i-1} \alpha_j\right) * (1-\pi_i) - \frac{\pi_i^2 + (1-\pi_i)^2}{2} \\ &= \pi_i * \left(\sum_{j=i}^{q} \alpha_j - \left(1 - \sum_{j=i}^{q} \alpha_j\right)\right) + \sum_{j=1}^{i-1} \alpha_j - \frac{2\pi_i^2 + 1 - 2\pi_i}{2}\end{aligned}$$

$$= 2 * \pi_i * \left(\sum_{j=i}^{q} \alpha_j\right) - \pi_i^2 + \frac{1}{2} + \sum_{j=1}^{i-1} \alpha_j$$
$$= 2 * \pi_i * \left(\sum_{j=i}^{q} \alpha_j\right) - \pi_i^2 + \frac{3}{2} - \sum_{j=i}^{q} \alpha_j$$

We obtain an evaluation π as a cumulated function by summing this calculus for all the π_i:

$$\begin{aligned}\mathcal{I}_{surf}(cumul(\pi)|X) &= \sum_i \mathcal{I}_{surf}((\pi_i, 1-\pi_i)|X) \\ &= 2 * \sum_{i=1}^{q} \left(\pi_i * \left(\sum_{j=i}^{q} \alpha_j\right)\right) - \sum_{i=1}^{q} \pi_i^2 + \sum_{i=1}^{q} \frac{3}{2} - \sum_{i=1}^{q} \left(\sum_{j=i}^{q} \alpha_j\right) \\ &= 2 * \sum_{i=1}^{q} \left(\pi_i * \left(\sum_{j=i}^{q} \alpha_j\right)\right) - \sum_{i=1}^{q} \pi_i^2 + \frac{3 * q}{2} - \sum_{i=1}^{q} \alpha_i * i\end{aligned} \tag{7}$$

Proposition 3. *Given a set of realizations $X = \{x_1, \ldots, x_n\}$, $\mathcal{I}_{surf}(cumul(\pi)|X)$ is equal to the same value for any π that is a cumulated function of p^**

Proof. (sketch) By replacing π by a cumulated function of p^* we obtain an expression that only depends on α_i's and that is independent of the ordering of the α_i's. □

Proposition 4. *Given a set of realizations $X = \{x_1, \ldots, x_n\}$ of a random variable over a discrete universe $\Omega = \{C_1, \ldots, C_q\}$ and given a fixed order for the π_i's finding the possibility distribution π that maximizes $\mathcal{I}_{surf}(cumul(\pi)|X)$ is equivalent to find the distribution that minimizes the square of the distance to the cumulated functions of p^* which have the α_i's and the π_i's in the same order.*

Proof. Given $X = \{x_1, \ldots, x_n\}$, the cumulated functions of p^* which have the α_i's and the π_i's in the same order is $\pi'_i = \sum_{j=i}^{q} \alpha_j$

$$\begin{aligned}d(\pi', \pi)^2 &= \sum_{i=1}^{q} \left(\sum_{j=i}^{q} \alpha_j - \pi_i\right)^2 \\ &= \sum_{i=1}^{q} \left(\left(\sum_{j=i}^{q} \alpha_j\right)^2 + \pi_i^2 - 2 * \left(\sum_{j=i}^{q} \alpha_j\right) * \pi_i\right) \\ &= \sum_{i=1}^{q} \left(\left(1 - \sum_{j=1}^{i-1} \alpha_j\right)^2 + \pi_i^2 - 2 * \left(\sum_{j=i}^{q} \alpha_j\right) * \pi_i\right) \\ &= \sum_{i=1}^{q} \left(1 + \left(\sum_{j=1}^{i-1} \alpha_j\right)^2 - \sum_{j=1}^{i-1} \alpha_j - 2 * \left(\sum_{j=i}^{q} \alpha_j\right) * \pi_i\right) + \pi_i^2 - \sum_{j=1}^{i-1} \alpha_j \\ &= C_c - \mathcal{I}_{surf}(cumul(\pi)|X)\end{aligned}$$

where $C_c = \sum_{i=1}^{q} \left(1 + \frac{3}{2} + \left(\sum_{j=1}^{i-1} \alpha_j\right)^2 - \sum_{j=1}^{i-1} \alpha_j\right)$ is independent of π. This establishes the proposition. □

Then $\mathcal{I}_{surf}(cumul(\pi)|X)$ is a good candidate for the possibilistic information distance since it is maximized by possibility distributions that are cumulated functions of p^*, and it can be still considered as the squared distance between the considered distribution and a cumulated function of p^*. However, Proposition 3 states that $\mathcal{I}_{surf}(cumul(\pi)|X)$ has the same value for any cumulated function. Thus, maximizing $\mathcal{I}_{surf}(cumul(\pi)|X)$ doesn't not guarantee that the α_i's and the π_i's are in the same order. In order to overcome this issue, we propose the following function:

$$\mathcal{I}_{pos}(\pi|X) = \sum_{i=1}^{q} \pi_i * \left(\sum_{j=i}^{q} \alpha_j\right) - \frac{1}{2} \sum_{i=1}^{q} \pi_i^2 - \sum_{i=1}^{q} \alpha_i * i \tag{8}$$

Indeed, we have:

$$\mathcal{I}_{pos}(\pi|X) = \frac{1}{2} * \mathcal{I}_{surf}(cumul(\pi)|X) - \frac{1}{2}\sum_{i=1}^{q} \alpha_i * i - \frac{3*q}{4}$$

$\mathcal{I}_{pos}$ is nothing but $\mathcal{I}_{surf}$ of the cumulated function with an additional term that favors the possibility distributions that respect the frequency ordering. This is established in the following proposition.

Proposition 5. *Given a set of data $X = \{x_1, \ldots, x_n\}$ belonging to a discrete universe $\Omega = \{C_1, \ldots, C_q\}$, the possibility distribution π^* that maximizes the function $\mathcal{I}_{pos}$ is the probability–possibility transformation of the optimal probability distribution p^* (i.e., $\forall i \in \{1, \ldots, q\}$, $p_i^* = \alpha_i$), according to Eq. (5) (i.e., $\pi^* = \pi^{sp}$).*

Proof. We look for the probability distribution π^* that maximizes $\mathcal{I}_{pos}$.
We have:

$$\forall i = 1 \ldots q, \frac{\delta \mathcal{I}_{pos}(\pi|X)}{\delta \pi_i} = \sum_{j=i}^{q} \alpha_j - \pi_i$$

thus

$$\forall i = 1 \ldots q, \frac{\delta \mathcal{I}_{pos}(\pi|X)}{\delta \pi_i} = 0 \Leftrightarrow p_i = \sum_{j=i}^{q} \alpha_j.$$

Since the derivative of $\mathcal{I}_{pos}(\pi|X)$ with respect to π_i (the parameters of π) does not depend on the other operator $\pi_j, j \neq i$, we obtain $\pi_i^* = p_i = \sum_{j=i}^{q} \alpha_j$ which corresponds to a cumulated distribution of the α_i's. Since the part $\sum_{i=1}^{q} \alpha_i * i$ is maximized when $\alpha_1 \geq \cdots \geq \alpha_q$, the distribution π^* corresponds exactly to Eq. (5) and we have $\pi^* = \pi^{sp}$. □

This proposition shows that $\mathcal{I}_{pos}$ is an acceptable informational distance function for possibility distributions viewed as families of probabilities. This proposition and Proposition 4 provide an interpretation of the possibility informational distance: if the maximization algorithm used does not prevent to have a possibility distribution that respect the frequency ordering, the algorithm will favor the distribution that shares the maximum of area with the optimal one. As for $\mathcal{I}_{surf}$ the informational distance depends on the surface shared between the considered possibility distribution and the optimal one.

If we only consider one piece of data x such that $x \in C_j$ we obtain :

$$\mathcal{I}_{pos}(\pi|x) = \frac{1}{n}\left(\sum_{i=1}^{j} \pi_i - \frac{1}{2}\sum_{i=1}^{q} \pi_i^2 - j\right) \tag{9}$$

It is worth noticing that, when optimal distributions can only be approximated, finding the best approximation with respect to $\mathcal{I}_{pos}$ is not equivalent to finding the best probability approximation with respect to a probabilistic informational distance and then turning it into a possibility distribution.

Example 2. For instance, we consider X that leads to the frequency $\alpha_1 = 0.5, \alpha_2 = 0.3, \alpha_3 = 0.2$. We require that $p_3 = 0$ and $\pi_3 = 0$. In this context, the optimal p with respect to $\mathcal{I}_{surf}$ ($\mathcal{I}_{log}$ is not applicable here) is $p_1 = 0.6, p_2 = 0.4, p_3 = 0$. The optimal π with respect to $\mathcal{I}_{pos}$ is $\pi_1 = 1, \pi_2 = 0.5, \pi_3 = 0$. The transformation π' of p is $\pi'_1 = 1, \pi'_2 = 0.4, \pi'_3 = 0$. We observe that π' is different than π and that π is a better approximation of the optimal possibility distribution given in Example 1.

This result is fundamental since it illustrates that using a probabilistic informational distance and then the probability–possibility transformation is not an effective approach for constructing a possibility distribution from data. The maximization of $\mathcal{I}_{pos}$ is more adapted in this scope. However, if the optimization constraints or the representation bias prevent us to have a possibility distribution that respect the frequency ordering, we may obtain the possibility distribution that shares the maximum of area with the cumulated distribution that corresponds to the ordering of the possibility degrees.

Example 3. We again consider X that leads to the frequency $\alpha_1 = 0.5, \alpha_2 = 0.3, \alpha_3 = 0.2$. We now require that $p_2 = 0$ and $\pi_2 = 0$. In this context, the optimal p with respect to $\mathcal{I}_{surf}$ ($\mathcal{I}_{log}$ is not applicable here) is $p_1 = 0.65, p_2 = 0, p_3 = 0.35$. The optimal π with respect to $\mathcal{I}_{pos}$ is $\pi_1 = 1, \pi_2 = 0, \pi_3 = 0.5$ which corresponds to the approximation of another distribution.

In this paper, the optimization of $\mathcal{I}_{pos}$ is performed with a simulated annealing algorithm [9]. It is interesting to remark that we have not restricted the state space to normalized distributions since the maximization of $\mathcal{I}_{pos}$ converges naturally to a normalized possibility distribution. However, the function $\mathcal{I}_{pos}$ also applies to non-normalized distributions, but since it evaluates the possibility distribution as a cumulated distribution function, normalized ones are always preferred. Then, if the optimization process returns a non-normalized distribution, normalizing it by fixing the greatest value to 1 will automatically increase the value of $\mathcal{I}_{pos}$.

5. Possibilistic informational distance: continuous case

We now extend the definition of the possibilistic informational distance to the continuous case where $\Omega = \mathbb{R}$.

5.1. Definitions

In the continuous case, the consideration of the values of π in an increasing order is naturally replaced by the consideration of α-cuts. Then the part $\sum_{i=1}^{j}(\pi_i) - j$ of Eq. (9) becomes:

$$\int_{A_{\pi(x)}} \pi(t)dt - |A_{\pi(x)}|$$

where $A_{\pi(x)}$ is the $\pi(x)$-cut of π and $|A_{\pi(x)}|$ its size. Naturally, the part $-\frac{1}{2}\sum_{i=1}^{q}\pi_i^2$ becomes:

$$-\frac{1}{2}\int_{\mathbb{R}} \pi(t)^2 dt$$

Then, for one piece of data we obtain:

$$\mathcal{I}_{pos}(\pi|x) = \int_{A_{\pi(x)}} \pi(t)dt - |A_{\pi(x)}| - \frac{1}{2}\int_{\mathbb{R}} \pi(t)^2 dt \tag{10}$$

If we consider more than one piece of data, we obtain:

$$\mathcal{I}_{pos}(\pi|X) = \frac{1}{n}\sum_{i=1}^{n}\left(\int_{A_{\pi(x_i)}} \pi(t)dt - |A_{\pi(x_i)}|\right) - \int_{\mathbb{R}} \frac{\pi(t)^2}{2} dt \tag{11}$$

Proposition 6. *Assuming an infinite set of data that follows a probability distribution p, finding the possibility distribution π that maximizes $\mathcal{I}_{pos}$ is equivalent to find the distribution that minimizes the square of the distance to the optimal one π^* if π respects the ordering of p (i.e., $\forall x, y \in \mathbb{R}$ if $p(x) \leq p(y)$ then $\pi(x) \leq \pi(y)$).*

Proof. It is basically the extension of Proposition 4 where the frequency becomes density values of p. Then, we will have:

$$d(\pi^*, \pi)^2 = C - 2 * \mathcal{I}_{pos}(\pi|p) + 2 * \int_{\mathbb{R}} p(x) * |A_{\pi(x)}|dx$$

where $C = \int_{\mathbb{R}}(\pi^*(x)^2 dx$ is independent of π. $2 * \int_{\mathbb{R}} p(x) * |A_{\pi(x)}|dx$ is minimized when values x with greater density are in the smallest α-cuts, i.e., when values x with greater density have greater possibility values. When π respects the ordering of p we have $A_{\pi(x)} = \{y \in \mathbb{R} | p(y) \geq p(x)\}$ which corresponds to the case when $2 * \int_{\mathbb{R}} p(x) * |A_{\pi(x)}|dx$ is minimal. □

Proposition 7. *The possibility distribution π^* that maximizes the function $\mathcal{I}_{pos}$ for data that follow a probability distribution p is the probability–possibility transformation of the optimal probability distribution p^* according to Eq. (4).*

Proof. It is the extension of Proposition 5 where the frequency becomes density values of p. We have:

$$\mathcal{I}_{pos}(\pi|p) = \int_{\mathbb{R}} p(x)\left(\int_{A_{\pi(x)}} \pi(t)dt - |A_{\pi(x)}|\right)dx - \int_{\mathbb{R}} \frac{\pi(t)^2}{2} dt \tag{12}$$

the part $\int_{\mathbb{R}} p(x)\left(\int_{A_{\pi(x)}} \pi(t)dt\right)dx - \int_{\mathbb{R}} \frac{\pi(t)^2}{2} dt$ is maximized for possibility distribution that are cumulated distribution of p (i.e., $A_{1-\alpha} = I_\alpha$). The part $-\int_{\mathbb{R}} p(x) * |A_{\pi(x)}|dx$ is maximized when values x with greater density are in the smallest α-cuts, i.e., when values x with greater density have greater density value, thus when π respects the ordering of p. Finally, $\mathcal{I}_{pos}$ is maximized when $A_{1-\alpha} = I_\alpha^*$, i.e., when $\pi^* = \pi^{sp}$. □

Thus, the possibilistic informational distance in the continuous case has the same understanding than in the discrete case. It favors the possibility distribution that minimizes the distance with the probability–possibility transformation of the distribution. However, the main advantage of this measure if that, of course, the optimal probability–possibility transformation has not to be known. Note that, having the informational distance function defined for triangular (resp. trapezoidal) distribution, we can obtain the optimal triangular (resp. trapezoidal) by finding the parameters that maximizes $\mathcal{I}_{pos}$. Since this problem cannot be solved analytically, we have to use a meta heuristics such as simulated annealing [9] or particle swarm optimization [8]. We will consider the calculus of $\mathcal{I}_{pos}$ for triangular and trapezoidal possibility distributions.

5.2. Triangular distribution

We define a triangular possibility distribution as the triple $\pi_{tri} = (m, l, r)$ where m is the mode of the triangle and l and r the left and the right spread respectively. We consider a piece of data $x \in X$. We note $\mu = \pi_{tri}(x)$ the possibility degree of x and $[a, b]$ the μ-cut of π_{tri}. The second part of the Eq. (10) is computed such as:

$$-\int_{\mathbb{R}} \frac{(\pi_{tri}(t))^2}{2} dt = -\int_{m-l}^{m+r} \frac{(\pi_{tri}(t))^2}{2} = -\frac{l+r}{6}.$$

There are two cases for the term that depends on $\pi_{tri}(x)$ in (10). We consider the case of $x \in]m-l, m+r[$. We have:

$$\begin{aligned}
\int_{A_\mu} \pi_{tri}(t)dt - |A_\mu| &= \int_a^b \pi_{tri}(t)dt - (1-\mu) * (l+r) \\
&= \int_a^m (\pi_{tri}(t))dt + \int_m^b (\pi_{tri}(t))dt - (1-\mu) * (l+r) \\
&= \int_a^m \left(1 - \frac{m-t}{l}\right) dt + \int_m^b \left(1 - \frac{t-m}{r}\right) dt - (1-\mu) * (l+r) \\
&= -(1-\mu)^2 * \frac{l+r}{2}.
\end{aligned}$$

The problem is when $x \notin]m-l, m+r[$. In this case, the 0-cut is infinite. In the settings of the proposition 7, the interval that has to be considered is $\{y \in \mathbb{R} | p(y) \geq p(x)\}$. If we assume that the probability decreases linearly when the distance to the bound increases, we obtain:

$$\int_{A_\mu} \pi_{tri}(t)dt - |A_\mu| = -\frac{l+r}{2} - C_1 * min(d(x, m-l), d(x, m+r)).$$

where C_1 is a constant and d denotes the Euclidean distance. Finally, we obtain:

$$\mathcal{I}_{pos}(\pi_{tri}|x) = \begin{cases} -(1-\mu)^2 * \frac{l+r}{2} - \frac{l+r}{6} & \text{if } x \in x \in]m-l, m+r[\\ -\frac{l+r}{2} - C_1 * d(x, m-l) - \frac{l+r}{6} & \text{if } x \leq (m-l) \\ -\frac{l+r}{2} - C_1 * d(x, m+r) - \frac{l+r}{6} & \text{if } x \geq (m+r) \end{cases} \tag{13}$$

Proposition 8. *Given an infinite set of data that follows an unimodal symmetric probability distribution p, the triangular possibility distribution that maximizes $\mathcal{I}_{pos}$ with $C_1 = 2$ is the one that minimizes the square of the distance to the optimal one π^{sp}.*

Proof. If the distribution is unimodal and symmetric, any triangular possibility distribution which has a mode that corresponds to the mode of the probability distribution will respect the ordering of p on its support. Outside the support, the value $-\frac{l+r}{2} - 2 * min(d(x, m-l), d(x, m+r)$ is equal to $\frac{l+r}{2} - (l + r + 2 * min(d(x, m-l), d(x, m+r))$. $l + r + 2 * min(d(x, m-l), d(x, m+r))$ is the size of the interval $\{y | p(y) \geq p(x)\}$ due to the symmetry of p. Then, the property 6 applies. □

This proposition shows that, if the distribution is unimodal and symmetric, the triangular possibility distribution that maximizes $\mathcal{I}_{pos}$ is the one that shares the maximal surface with the probability–possibility transformation of the considered distribution. Moreover, it has been shown in [3] that, given an unimodal and symmetric probability distribution, for any value of α, it exists a triangular possibility distribution for which for all $\beta \leq \alpha$, $I^*_\beta \subseteq A_{1-\beta}$. This triangular distribution can also be obtained by maximizing $\mathcal{I}_{pos}$ with the adequate value of C_1. In fact, when C_1 increases, the weight of the value outside the support increases and then the size of the support of the optimal distribution will increase too. Finally, the maximization of $\mathcal{I}_{pos}$ for a triangular possibility distributions can be used for upper estimating the confidence intervals of an unimodal and symmetric probability distribution, until a confidence level threshold. This threshold increases when C_1 increases.

This result does not hold when considering asymmetric unimodal probability distribution. This is essentially due to the fact that the confidence intervals, and then the α-cuts are not centered on the mean of the distribution (as it is considered to be for values outside the support in Eq. (13)). We propose to adapt Eq. (13) in order to take into account this asymmetry in the following way:

$$\mathcal{I}_{pos}(\pi_{tri}|x) = \begin{cases} -(1-\mu)^2 * \frac{l+r}{2} - \frac{l+r}{6} & \text{if } x \in x \in]m-l, m+r[\\ -\frac{l+r}{2} - \frac{1}{2} * C_1 * \left(1 + \frac{r}{l}\right) * d(x, m-l) - \frac{l+r}{6} & \text{if } x \leq (m-l) \\ -\frac{l+r}{2} - \frac{1}{2} * C_1 * \left(1 + \frac{l}{r}\right) * d(x, m+r) - \frac{l+r}{6} & \text{if } x \geq (m+r) \end{cases} \tag{14}$$

Even if this version of $\mathcal{I}_{pos}$ does not guarantee that the results obtained for the symmetric unimodal probability distribution remain true for asymmetric unimodal ones, we observe that it provides good approximations in most of the cases. Indeed, the optimal triangular possibility distribution with respect to $\mathcal{I}_{pos}$ is the approximation of the cumulated distribution that respect as much as possible the ordering of the density function ($p(x) \geq p(x') \leftrightarrow \pi(x) \geq \pi(x')$). Triangular distribution can respect this ordering in the symmetric case, it is not always the case in the non symmetric case. If the probability values are proportional to the distance of the mode (i.e., $p(x+m) = c * p(x-m)$ where m is the mode and c a constant) the mode of optimal triangular possibility distribution will correspond to the mode of the probability distribution. The less the probability values are proportional to the distance of the mode, the higher the distance between the mode of optimal triangular possibility distribution and the mode of the probability distribution.

5.3. Trapezoidal distribution

We define a trapezoidal possibility distribution as the quadruple $\pi_{trap} = (a, b, c, d)$ where $[a, d]$ and $[b, c]$ are respectively the support and the core of the distribution. We consider a piece of data $x \in X$. We note $\mu = \pi_{trap}(x)$ the possibility degree of x and $[a, b]$ the μ-cut of π_{trap}. We also note $lt = (b-a)$, the size of the left part of the support, $rt = (d-c)$ the size of the right part of the support and $mt = (c-b)$ the size of the core. In the spirit of the triangular case we obtain:

$$\mathcal{I}_{pos}(\pi_{trap}|x) = \begin{cases} -(1-\mu)^2 * \frac{lt+rt}{2} - \frac{lt+rt}{6} - C_2 * mt & \text{if } x \in x \in]b, c[\\ -\frac{lt+rt}{2} - \frac{1}{2} * C_1 * \left(1 + \frac{rt}{lt}\right) * d(x, a) - \frac{lt+rt}{6} - C_2 * mt & \text{if } x \leq a \\ -\frac{lt+rt}{2} - \frac{1}{2} * C_1 * \left(1 + \frac{lt}{rt}\right) * d(x, d) - \frac{lt+rt}{6} - C_2 * mt & \text{if } x \geq d \end{cases} \tag{15}$$

where C_2 is a constant. Formally, C_2 is equal to 1. However, in such a case, the optimal distribution with respect to $\mathcal{I}_{pos}$ will always be a triangular one (i.e., $b = c$). This is due to the fact that increasing the core of the distribution will never decrease the distance with the optimal distribution. By allowing to have C_2 less than 1, we allow the decreasing of the weight of the core of the distribution and then to have a genuine trapezoidal possibility distribution that maximizes $\mathcal{I}_{pos}$. Notice that it is not in contradiction with the spirit of the possibilistic informational distance since the optimal trapezoidal distribution obtained by decreasing C_2 will always contain the optimal triangular one (for which $C_2 = 1$). C_1 has the same use and meaning than in the triangular case. Finally, the maximization of $\mathcal{I}_{pos}$ for a trapezoidal possibility distribution with adequate values of C_1 and C_2, can be used for upper estimating the confidence intervals of any multi modal (with a finite number of mode) probability distribution, until a confidence level threshold. As in the triangular case, this threshold increases when C_1 increases. Unlike the discrete case, the use of triangular distribution guarantee by construction that the distribution obtained by optimization of $\mathcal{I}_{pos}$ is normalized even the definition of $\mathcal{I}_{pos}$ is still relevant if the distribution is not normalized.

6. Illustrations

In this section, we use the possibilistic informational distance function in order to a build possibility distribution from a set of data. In each case, we will consider a different probability distribution. Thus, we construct a set of data that corresponds as much as possible to the probability distribution chosen. In order to do that, we discretize the density function on a fixed interval. After a normalization step, we obtain a discrete probability distribution. We then maximize the sum of the $\mathcal{I}_{pos}$ for each of these points weighted by their probability degree. We choose this approach rather than a sampling because, with a sufficient high discretization range, the result obtained is very close to the optimization of $\mathcal{I}_{pos}$ for an infinite set of data that follows the chosen probability distribution. Of course, the method is still applicable to any sample set. As pointed out previously, we have to use a meta heuristic (simulated annealing here) in order to find the maximal triangular or trapezoidal distribution that maximize $\mathcal{I}_{pos}$. For each probability distribution, we discretize the density function with 500 divisions of the interval $[-6, 6]$. On each figure, the probability distribution is in green, its probability–possibility transformation is in blue

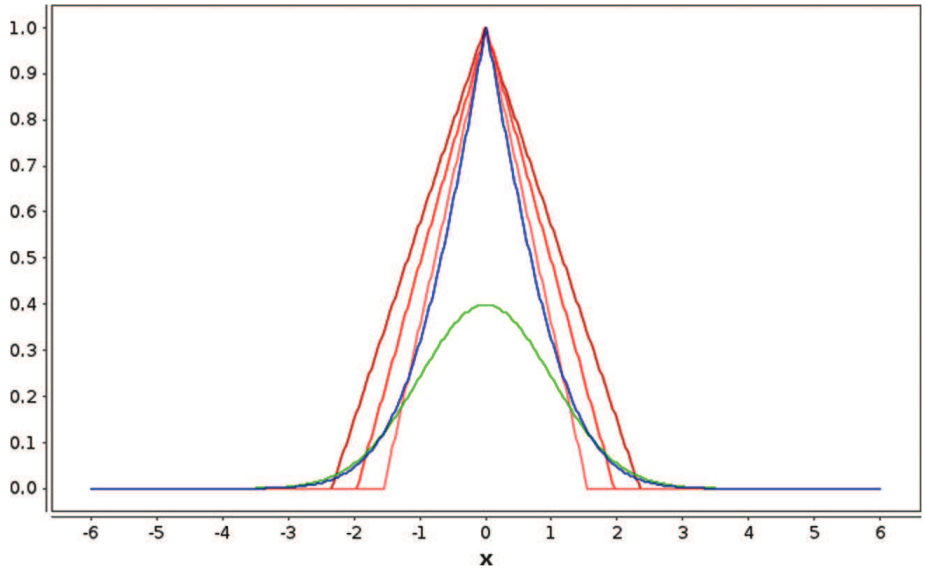

Fig. 3. Triangular possibility distributions that maximize $\mathcal{I}_{pos}$ with respect to a set of data that follows a Gaussian distribution.

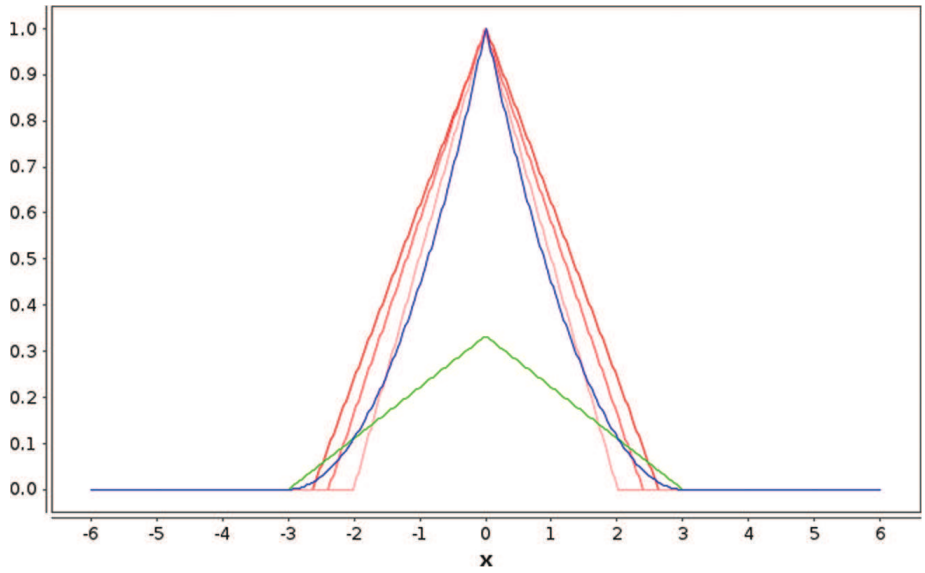

Fig. 4. Triangular possibility distributions that maximize $\mathcal{I}_{pos}$ with respect to a set of data that follows a triangular probability distribution.

and optimal possibility distributions are in red. A figure may contain multiple possibility distributions which correspond to different values of the constants C_1 and C_2.

In the following, for illustration purpose we suppose that we know the shape of the probability distribution. Obviously, in this case, the squared distance with the probability–possibility transformation can be computed directly. However, $\mathcal{I}_{pos}$ does not require the knowledge of the shape of the distribution. Moreover, the result is obtained by summing up the local evaluations of $\mathcal{I}_{pos}$ on each piece of data, while the method based on the square of the surface relies on a global evaluation. Contrarily to the method based on the square of the surface, the optimization of $\mathcal{I}_{pos}$ still applies when having a finite (even small) amount of data, without knowing the shape of the distribution.

6.1. Unimodal symmetric distributions

First, we consider a unimodal symmetric distribution. Fig. 3 corresponds to the building of triangular possibility distribution that maximize $\mathcal{I}_{pos}$ with $C_1 = 2$, $C_1 = 4$ and $C_1 = 10$ (from light red to dark red)[1] for data that follow a Gaussian distribution with mean equal to 0 and standard deviation equal to 1. We can first notice that the mode of the triangular distribution corresponds to the mode of the Gaussian distribution. The triangular distribution with $C_1 = 2$ is very close to the optimal one. As expected, when C_1 increases, the threshold α for which $\forall\beta \leq \alpha, I^*_\beta \subseteq A_{1-\beta}$ increases too. In the

[1] For interpretation of colour in figure artwork, the reader is referred to the web version of this article.

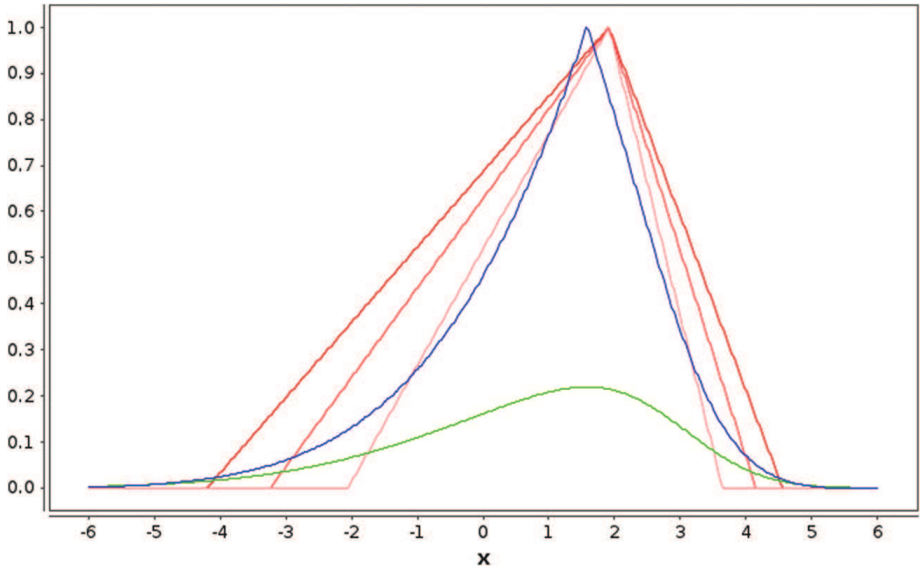

Fig. 5. Triangular possibility distributions that maximize $\mathcal{I}_{pos}$ with respect to a set of data that follows a skewed normal probability distribution.

following, we will name domination level the value α. Of course we have $\pi_{C_1=2} \preceq \pi_{C_1=4} \preceq \pi_{C_1=10}$. For $\pi_{C_1=2}$, $\pi_{C_1=4}$ and $\pi_{C_1=10}$, the domination levels are respectively 0.76, 0.93 and 0.97.

Fig. 4 presents the same approach with data that follow a triangular probability distribution. The results are similar to the ones observed with a Gaussian distribution. The domination levels are very close (respectively 0.76, 0.93 and 0.97). As for the discrete case, this figure illustrates the fact that maximizing possibilistic informational distance provides results that are different from the ones obtained by approximating the probability distribution and then applying the probability–possibility transformation. This is obvious here since the probability–possibility transformation of a triangular probability distribution is not a triangular possibility distribution.

What is interesting to remark here is that the maximum possibilistic informational distance principle allows us to elicitate triangular possibility distributions that upper bound the confidence intervals of any unknown unimodal symmetric probability distribution.

6.2. Unimodal non symmetric distributions

We perform the same experience than the previous one with a skewed normal distribution which is unimodal and asymmetric (see Fig. 5). Although the distribution is not symmetric, we can observe similar results to the ones observed with the Gaussian distribution, except that the mode is not identified exactly. Thus, $\pi_{C_1=2}$ upper bounds the confidence intervals I^*_α for $0.2 \leq \alpha \leq 0.76$, $\pi_{C_1=4}$ for $0.1 \leq \alpha \leq 0.94$ and $\pi_{C_1=10}$ for $0.07 \leq \alpha \leq 0.98$.

We can observe that, even if the results are not as precise as the previous ones, maximum possibilistic informational distance principle is still able to upper bound unknown unimodal distributions that are not symmetric. We can also observe that, even the choice of C_1 is empirical at this time, the dominance level remains relatively stable for a given value of C_1, regardless to the type of the distribution.

6.3. Multimodal symmetric distributions

In Figs. 6–8, the distribution is multi modal. In Fig. 6 we have $C_2 = 1$, and from light red to dark red $C_1 = 2$, $C_1 = 4$ and $C_1 = 10$. As expected, we obtain, we obtain triangular distributions. We can observe that the mode of the triangular distribution is close the highest mode of the probability distribution. When $C_2 = 0.8$ (Fig. 7), we obtain genuine trapezoidal distributions, but they still do not upperbound the confidence intervals. A possibility distribution that upper bounds the confidence intervals I^*_α for $\alpha \leq 0.98$ is obtained with $C_1 = 10$ and $C_2 = 0.67$ (Fig. 8). This illustrates the fact that, for adequate values of C_1 and C_2, the maximum possibilistic informational distance principle allows us to elicitate trapezoidal possibility distributions that upper bound the confidence intervals of any unknown probability distribution that has a finite number of modes. As for C_1, the effect of the value of C_2 does not highly depend on the type of the distribution and then the values $C_1 = 10$ and $C_2 = 0.67$ perform well in most of the cases.

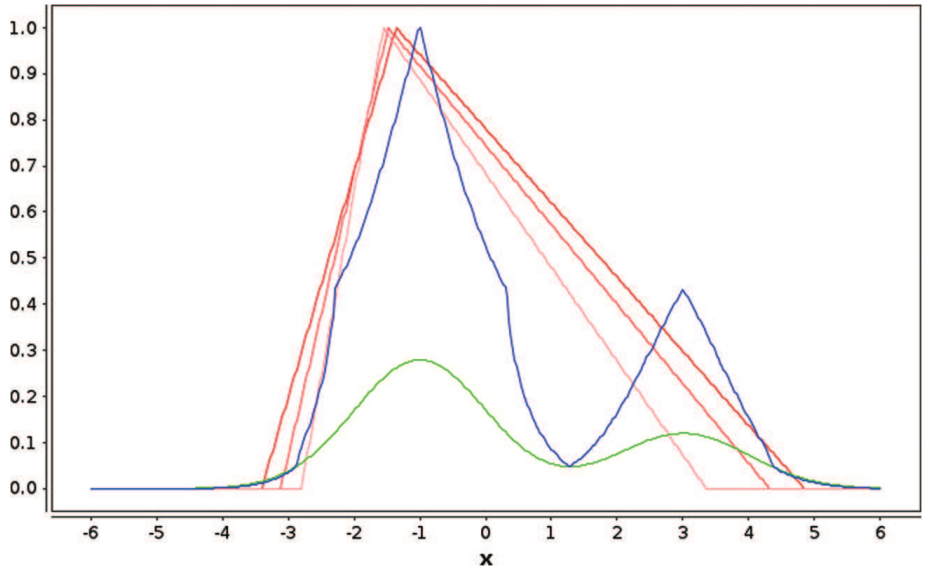

Fig. 6. Trapezoidal possibility distributions that maximize $\mathcal{I}_{pos}$ ($C_2 = 1$) with respect to a set of data that follows a Gaussian probability distribution.

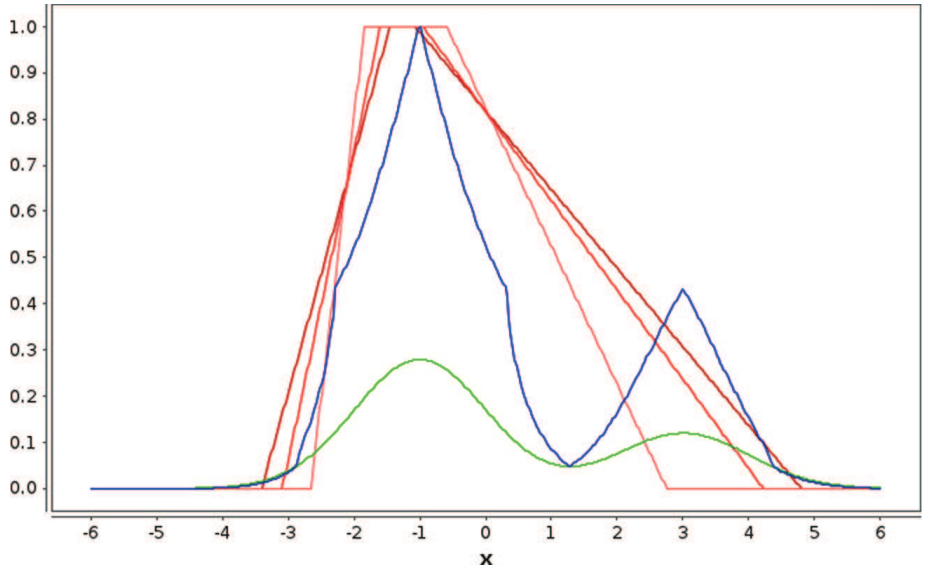

Fig. 7. Trapezoidal possibility distributions that maximize $\mathcal{I}_{pos}$ ($C_2 = 0.8$) with respect to a set of data that follows a Gaussian probability distribution.

7. Conclusion

In this paper we have proposed a definition of possibilistic informational distance that agrees with the view of possibility distributions as families of probability distributions and with the probability–possibility transformation based on the maximum-specificity principle. We have defined this possibilistic informational distance function both for discrete and for continuous universes. The calculus is described for the cases of triangular and trapezoidal possibility distributions. We have shown that the possibilistic informational distance is related to the distance between the considered possibility distribution and the probability–possibility transformation of the unknown probability distribution that has generated the data. More precisely, by maximizing $\mathcal{I}_{pos}$ (under some constraints, such as triangular shaped possibility distributions) we obtain the possibility distribution that respects as much as possible the ordering of the probabilities, and that shares a maximal surface with the cumulated distribution that respects the possibilistic ordering.

This type of function is interesting in many respects. It can be used for comparing the faithfulness of two possibility distributions with respect to a set of data. Moreover, the good properties of triangular possibility distributions for bounding the confidence intervals of a unimodal probability distribution makes the building of possibility distributions by optimization of $\mathcal{I}_{pos}$ a good approach when no a priori information on the type of distribution is available. In the same way, we can use maximum possibilistic informational distance principle for upper bounding the confidence intervals of any unknown probability distribution that has a finite number of modes.

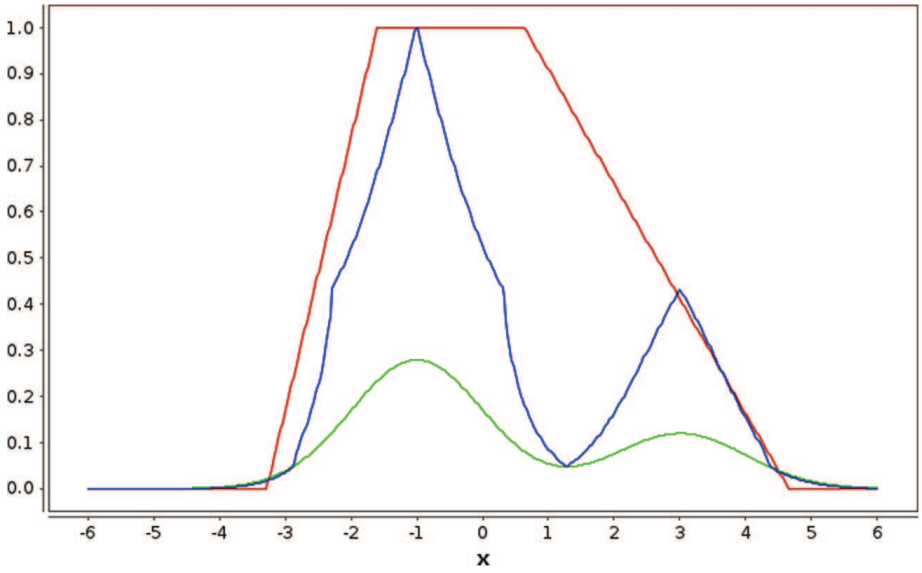

Fig. 8. Trapezoidal possibility distribution that maximizes $\mathcal{I}_{pos}$ ($C_1 = 10, C_2 = 0.67$) with respect to a set of data that follows a Gaussian probability distribution.

Possibilistic informational distance is particularly promising in the area of machine learning. Indeed, it is common in machine learning to have to estimate probability distributions with few data and without a priori knowledge about the shape of the distribution. It may happen for instance in Bayesian learning or in k-nearest neighbor approaches. In this case, possibilistic informational distance may be a cautious and valuable tool for upper bounding the confidence intervals of these unknown distributions. It this scope, maximizing the possibilistic informational distance is the core of the imprecise regression method [14] that allows us to predict a possibility distribution of the output value from a crisp vector of input values. A similar approach to possibilistic classification is worth investigating too.

In the future, it would be useful to look for alternatives to the meta heuristics that are used in the optimization process since these algorithms are heavy to tune and require high amounts of computation. This may rely on analytical solutions of the optimization problems, on the use of lighter algorithms that provide good approximations. Moreover, another practical issue is the tuning of the parameters C_1 and C_2 and the study of their properties. Lastly, it may desirable to take into account the quantity of data available. Indeed, if only one value is available, the best possibility distribution is a Dirac, and there will be no difference with the case where a thousand of identical values are available. Less data should lead to less specific distributions. This should be taken into account in the computation of the possibilistic informational distance.

References

[1] A. Aregui, T. Denoeux, Constructing consonant belief functions from sample data using confidence sets of pignistic probabilities, International Conference on Pattern Recognition 49 (2008) 575–594.
[2] D. Dubois, Possibility theory and statistical reasoning, Computational Statistics and Data Analysis 51 (2006) 47–69.
[3] D. Dubois, L. Foulloy, G. Mauris, H. Prade, Probability–possibility transformations, triangular fuzzy sets, and probabilistic inequalities, Reliable Computing 10 (2004) 273–297.
[4] D. Dubois, H. Prade, When upper probabilities are possibility measures, Fuzzy Sets and Systems 49 (1992) 65–74.
[5] D. Dubois, H. Prade, On data summarization with fuzzy sets, in: Proceedings of the 5th International Fuzzy Systems Association World Congress (IFSA'93), Seoul, 1993, pp. 465–468.
[6] D. Dubois, H. Prade, Possibility theory: qualitative and quantitative aspects, in: D.M. Gabbay, Ph. Smets (Eds.), Quantified Representation of Uncertainty and Imprecision, Handbook of Defeasible Reasoning and Uncertainty Management Systems, vol. 1, Kluwer Academic Publishers., 1998, pp. 169–226.
[7] D. Dubois, H. Prade, S. Sandri, On possibility/probability transformations, Proceedings of Fourth IFSA Conference, Kluwer Academic Publ., 1993, pp. 103–112.
[8] J. Kennedy, R. Eberhart, Particle swarm optimization, in: Neural Networks, 1995. Proceedings, IEEE International Conference on, 1995, pp. 1942–1948.
[9] S. Kirkpatrick, C.D. Gelatt, M.P. Vecchi, Optimization by simulated annealing, Science 220 (1983) 671–680.
[10] K. Loquin, O. Strauss, Fuzzy histograms and density estimation, in: Soft Methods for Integrated Uncertainty Modelling, Proceedings of the 2006 International Workshop on Soft Methods in Probability and Statistics, SMPS 2006, Bristol, UK, 5–7 September 2006.
[11] M.-H. Masson, T. Denoeux, Inferring a possibility distribution from empirical data, Fuzzy Sets and Systems 157 (2006) 319–340.
[12] G. Mauris, Inferring a possibility distribution from very few measurements, Soft Methods for Handling Variability and Imprecision, Advances in Soft Computing, vol. 48, Springer, Berlin / Heidelberg, 2008, pp. 92–99.
[13] G. Mauris, Possibility distributions: A unified representation of usual direct-probability-based parameter estimation methods, International Journal of Approximate Reasoning 52 (9) (2011) 1232–1242.
[14] M. Serrurier, H. Prade, Imprecise regression based on possibilistic likelihood, in: Scalable Uncertainty Management – 5th International Conference Proceedings, SUM 2011, Dayton, OH, USA, October 10–13, 2011, pp. 447–459.
[15] M. Serrurier, H. Prade, Maximum-likelihood principle for possibility distributions viewed as families of probabilities, in: FUZZ-IEEE 2011, IEEE International Conference on Fuzzy Systems, Taipei, Taiwan, 27–30 June, 2011, Proceedings, 2011, pp. 2987–2993.
[16] O. Strauss, F. Comby, M. Aldon, Rough histograms for robust statistics, International Conference on Pattern Recognition 2 (2000) 2684
[17] L.A. Zadeh, Fuzzy sets as a basis for a theory of possibility, Fuzzy Sets and Systems 1 (1978) 3–25.